AMENDMENTS TO THE CLAIMS

The listing of claims will replace all prior versions, and listings of claims in the application:

LISTING OF CLAIMS:

1. (Currently amended) A method for identifying output documents similar to an input document, comprising:

(a) receiving the input document that includes textual content

(b) performing optical character recognition on the textual content to identify text;

(c) analyzing the text and the textual content to identify keywords, wherein ~~identifying~~ a predefined number of keywords is identified from a first list of rated keywords extracted from the input document;

(d) creating a list of best keywords wherein for each keyword remaining in the first list of keywords performing the steps,

(1) identifying the keyword in one or more domain specific dictionaries of words and phrases in which they are used;

(2) identifying combinations of keywords in the list of keywords that satisfy the longest phrase;

(3) determining the frequency of occurrence in the input document of the identified keywords and phrases identified in the one or more domain specific dictionaries;

(4) setting the linguistic frequency of occurrence of the keywords and phrases to a predefined value; and

(e) defining a ~~to define a~~ list of best keywords, wherein the list of best keywords has a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency;

(f) ~~(b)~~ formulating a query using the list of best keywords;

(g) ~~(c)~~ performing the query to assemble a first set of output documents;

(h) ~~(d)~~ identifying lists of keywords for each output document in the first set of documents by tokenizing the keywords at one or more predefined word boundaries while

maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(i) ~~(e)~~ computing a measure of similarity between the input document and each output document in the first set of documents; and

(j) ~~(f)~~ defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value; wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency, each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document, wherein the query is repeated until a predetermined number of results are obtained or the query is terminated

(k) ~~(g)~~ if the second set of documents includes a matching document but no similar documents repeating (a)-(j)~~(f)~~ using the matching document to identify similar documents, wherein if one or more documents is related to a copyright registered document, the one or more documents is rights limited; and

(l) delivering each document in the second set of documents to one or more predetermined output devices.

2. (Cancelled)

3. (Currently amended) The method according to claim 1, further comprising

(m)~~(h)~~ if the second set of document contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency.

4. (Currently amended) The method according to claim 3, further comprising if after performing (m)~~(h)~~ the second set of document contains an insufficient number of output documents, performing

(n)~~(i)~~: replacing the list of best keywords using keywords having a rating greater than other keywords in the first list of rated keywords; and repeating (b)-(l)~~(f)~~.

5. (Cancelled)

6. (Currently amended) The method according to claim ~~4~~5, performing (n)~~(i)~~ when textual content in the input document is identified using OCR or a portion of the input document matches the output document.

7. (Currently amended) The method according to claim ~~5~~1, wherein the predefined number of keywords identified from the first list of rated keywords is five.

8. (Cancelled)

9. (Original) The method according to claim 1, further comprising:
recording a digital image representation of the input document;
performing OCR on the digital image representation to identify text;
analyzing the text to identify keywords.

10. (Currently amended) The method according to claim 1, further comprising:
(o)~~(k)~~ extracting from the input document the first list of keywords;
(p)~~(l)~~ determining if each keyword in the first list of keywords exists in a domain specific dictionary of words;
(q)~~(m)~~ for each keyword in the first list of keywords, determining its frequency of occurrence in the input document, also referred to as its term frequency;
(r)~~(n)~~ for each keyword identified at (k) that exists in the domain specific dictionary of words, assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies;

(s)~~(o)~~ for each keyword that was not identified in the domain specific dictionary of words at (h), assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; and

(t)~~(p)~~ for each keyword in the first list of keywords to which a term frequency and a linguistic frequency are assigned, computing a rating corresponding to its importance in the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence in the collection of documents.

11. (Currently amended) The method according to claim 10, for each keyword that was not identified in the domain specific dictionary of words at (p)~~(l)~~ and that was not assigned at (r)~~(n)~~ a linguistic frequency from the database of linguistic frequencies, assigning each that matches a regular expression from a set of regular expressions a predefined rating.

12. (Currently amended) The method according to claim 11, further comprising, for each keyword in the first list of keywords, modifying the term frequency of keywords determined at (q)~~(m)~~ to a predefined maximum.

13. (Original) The method according to claim 12, wherein keywords include phrases of keywords.

14. (Original) The method according to claim 11, wherein the rating is a weight computed using the following equation: $W_{t,d} = F_{t,d} * \log(N / F_t)$, where:

$W_{t,d}$: the weight of term t in document d;

$F_{t,d}$: the frequency occurrence of term t in document d;

N: the number of documents in the collection of documents;

F_t: the document linguistic frequency of term t in the collection of documents.

15. (Original) The method according to claim 11, wherein keywords that do not match a regular expression from the set of regular expressions are removed from the first list of keywords.

16. (Currently amended) A method for computing ratings of keywords extracted from an input document, comprising:

(a) determining if each keyword in the list of keywords exists in a domain specific dictionary of words by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(b) determining a frequency of occurrence in the input document for each keyword in the list of keywords, also referred to as its term frequency;

(c) for each keyword identified at (a) that exists in the domain specific dictionary of words, assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies;

(d) for each keyword that was not identified in the domain specific dictionary of words at (a), assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; and

(e) for each keyword in the list of keywords to which a term frequency and a linguistic frequency are assigned, computing a rating corresponding to its importance in the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence in the collection of documents, wherein a query reduction is performed by removing at least one keyword in the list of best keywords that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency if an insufficient number of results are obtained from the list of keywords, wherein the query is repeated until a predetermined number of results are obtained or the query is terminated ~~(f) if the second set of documents includes a matching document but no similar documents repeating (a)-(f) using the matching document to identify similar documents~~, wherein if one

or more documents is a copy of a known copyright registered document, the one or more documents is rights limited; and

(g) delivering each document in the collection of documents to a predetermined output device, wherein the collection of documents is set forth in a list serialized in XML that contains for each document found: its location on a network, original representation, unformatted representation, service results, metadata, distance measurement, type of document found according to desired quality, and error status.

17. (Original) The method according to claim 16, wherein the keywords in the list of keywords are used to carry out one of language identification, indexing, categorization, clustering, searching, translating, storing, duplicate detection, and filtering.

18. (Currently amended) A system for identifying output documents similar to an input document, comprising: a memory for storing the output documents and the input document and processing instructions of the system; and a processor coupled to the memory for executing the processing instructions of the system; the processor in executing the processing instructions:

(a) identifying a predefined number of keywords from a first list of rated keywords extracted from the input document ~~to define a list of best keywords~~;

(b) creating a list of best keywords wherein for each keyword remaining in the first list of keywords performing the steps,

(1) identifying the keyword in one or more domain specific dictionaries of words and phrases in which they are used;

(2) identifying combinations of keywords in the list of keywords that satisfy the longest phrase;

(3) determining the frequency of occurrence in the input document of the identified keywords and phrases identified in the one or more domain specific dictionaries;

(4) setting the linguistic frequency of occurrence of the keywords and phrases to a predefined value; and

(c) defining a list of best keywords wherein the list of best keywords have ~~having~~ a rating greater than other keywords in the first list of keywords except for keywords

belonging to a domain specific dictionary of words and having no measurable linguistic frequency by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(d) ~~(b)~~ formulating a query using the list of best keywords;

(e) ~~(c)~~ performing the query to assemble a first set of output documents;

(f) ~~(d)~~ identifying lists of keywords for each output document in the first set of documents;

(g) ~~(e)~~ computing a measure of similarity between the input document and each output document in the first set of documents;

(h) ~~(f)~~ defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value; wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency; and

(i) ~~(g)~~ if the second set of document contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency, wherein the query is repeated until a predetermined number of results are obtained or the query is terminated

(j) ~~(h)~~ if the second set of documents includes a matching document but no similar documents repeating (a)-(i)~~(g)~~ using the matching document to identify similar documents~~.~~, wherein if one or more documents is a copy of a known copyright registered document, the one or more documents is rights limited.

19. (Currently amended) The system according to claim 18, wherein the processor in executing the processing instructions further comprises:

~~(i)~~ (k) extracting from the input document the first list of keywords;

~~(j)~~ (l) determining if each keyword in the first list of keywords exists in a domain specific dictionary of words;

~~(k)~~(m) for each keyword in the first list of keywords, means for determining its frequency of occurrence in the input document, also referred to as its term frequency;

~~(l)~~ (n) for each keyword identified at ~~(j)~~ (l) that exists in the domain specific dictionary of words, means for assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies;

~~(m)~~ (o) for each keyword that was not identified in the domain specific dictionary of words at ~~(i)~~(l), means for assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; and

~~(n)~~(p) for each keyword in the first list of keywords to which a term frequency and a linguistic frequency are assigned, means for computing a rating corresponding to its importance in the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence in the collection of documents.

20. (Currently amended) An article of manufacture for identifying output documents similar to an input document, the article of manufacture comprising computer usable media including computer readable instructions embedded therein that causes a computer to perform a method, wherein the method comprises:

(a) identifying a predefined number of keywords from a first list of rated keywords extracted from the input document to define a list of best keywords; the list of best keywords having a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency, wherein the keywords are tokenized at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

~~(b) formulating a query using the list of best keywords;~~

(b) creating a list of best keywords wherein for each keyword remaining in the first list of keywords performing the steps,

(1) identifying the keyword in one or more domain specific dictionaries of words and phrases in which they are used;

(2) identifying combinations of keywords in the list of keywords that satisfy the longest phrase;

(3) determining the frequency of occurrence in the input document of the identified keywords and phrases identified in the one or more domain specific dictionaries;

(4) setting the linguistic frequency of occurrence of the keywords and phrases to a predefined value; and

(c) defining a list of best keywords wherein the list of best keywords have a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(d) formulating a query using the list of best keywords;

(e) ~~(c)~~ performing the query to assemble a first set of output documents;

(f) ~~(d)~~ identifying lists of keywords for each output document in the first set of documents;

(g) ~~(e)~~ computing a measure of similarity between the input document and each output document in the first set of documents;

(h) ~~(f)~~ defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value; wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency, each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document; and

(i) ~~(g)~~ if the second set of document contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency, wherein the query is repeated until a predetermined number of results are obtained or the query is terminated

(j) ~~(h)~~ if the second set of documents includes a matching document but no similar documents repeating (a)-(i)~~(g)~~ using the matching document to identify similar documents; and

(k) delivering each document in the second set of documents to a predetermined output device, wherein the collection of documents is set forth in a list serialized in XML that contains for each document found: its location on a network, original representation, unformatted representation, service results, metadata, distance measurement, type of document found according to desired quality, and error status.

21. (Currently amended) The system according to claim 18, further comprising if after performing ~~(g)~~(i) the second set of document contains an insufficient number of output documents, performing:

~~(i)~~(l) replacing the list of best keywords using keywords having a rating greater than other keywords in the first list of rated keywords; and repeating (b)-~~(f)~~(k).

22. (Currently amended) The system according to claim 18, wherein for each keyword that was not identified in the domain specific dictionary of words at ~~(f)~~(h) and that was not assigned at ~~(g)~~(i) a linguistic frequency from the database of linguistic frequencies, assigning each that matches a regular expression from a set of regular expressions a predefined rating, wherein the rating is a weight computed using the following equation: $W_{t,d} = F_{t,d} * \log(N / F_t)$, where:

$W_{t,d}$: the weight of term t in document d;

$F_{t,d}$: the frequency occurrence of term t in document d;

N: the number of documents in the collection of documents;

F_t: the document linguistic frequency of term t in the collection of documents.